Regime-Specific Predictability in Predictive Regressions

Jesús Gonzalo
Department of Economics, Universidad Carlos III de Madrid, Getafe (Madrid) 28903, Spain
(*jesus.gonzalo@uc3m.es*)

Jean-Yves Pitarakis
Economics Division, School of Social Sciences, University of Southampton, Southampton SO17 1BJ, UK
(*j.pitarakis@soton.ac.uk*)

Predictive regressions are linear specifications linking a noisy variable such as stock returns to past values of a very persistent regressor with the aim of assessing the presence of predictability. Key complications that arise are the potential presence of endogeneity and the poor adequacy of asymptotic approximations. In this article, we develop tests for uncovering the presence of predictability in such models when the strength or direction of predictability may alternate across different economically meaningful episodes. An empirical application reconsiders the dividend yield-based return predictability and documents a strong predictability that is countercyclical, occurring solely during bad economic times. This article has online supplementary materials.

KEY WORDS: Endogeneity; Persistence; Return predictability; Threshold models.

1. INTRODUCTION

Predictive regressions with a persistent regressor (e.g., dividend yields, interest rates, realized volatility) aim to uncover the ability of a slowly moving variable to predict future values of another typically noisier variable (e.g., stock returns, GDP growth) within a bivariate regression framework. Their pervasive nature in many areas of economics and finance and their importance in the empirical assessment of theoretical predictions of economic models made this particular modeling environment an important and active area of theoretical and applied research (for instance, see Jansson and Moreira 2006 and references therein).

A common assumption underlying old and new developments in this area involves working within a model in which the persistent regressor enters the predictive regression linearly, thus not allowing for the possibility that the strength and direction of predictability may themselves be a function of some economic factor or time itself. Given this restriction, existing work has focused on improving the quality of estimators and inferences in this environment, characterized by persistence and endogeneity, among other econometric complications. These complications manifest themselves in the form of nonstandard asymptotics, distributions that are not free of nuisance parameters, poor finite-sample approximations, etc. Important recent methodological breakthroughs have been obtained by Jansson and Moreira (2006), Campbell and Yogo (2006), Valkanov (2003), and Lewellen (2004), while recent applications in the area of financial economics and asset pricing can be found in Cochrane (2008), Lettau and Nieuwerburgh (2008), Bandi and Perron (2008), among others.

The purpose of this article is to instead develop an econometric toolkit for uncovering the presence of predictability within regression models with highly persistent regressors when the strength or direction of predictability, if present, may alternate across different economically meaningful episodes (e.g., periods of rapid versus slow growth, periods of high versus low stock market valuation, periods of high versus low consumer confidence, etc.). For this purpose, we propose to expand the traditional linear predictive regression framework to a more general environment that allows for the possibility that the strength of predictability may itself be affected by observable economic factors. We have in mind scenarios whereby the predictability induced by some economic variable kicks in under particular instances, such as when the magnitude of the variable in question (or some other variable) crosses a threshold but is useless in terms of predictive power otherwise. Alternatively, the predictive impact of a variable may alternate in sign/strength across different regimes. Ignoring such phenomena by proceeding within a linear framework, as has been done in the literature, may mask the forecasting ability of a particular variable and, more generally, mask the presence of interesting and economically meaningful dynamics. We subsequently apply our methodology to the prediction of stock returns with dividend yields (DY). Contrary to what has been documented in the linear predictability literature, our findings strongly point toward the presence of regimes in which DY-based predictability kicks in solely during bad economic times. More importantly, our analysis also illustrates the fact that the presence of regimes may make predictability appear as nonexistent when assessed within a linear model.

The plan of the article is as follows. Section 2 introduces our model and hypotheses of interest. Section 3 develops the limiting distribution theory of our test statistics. Section 4 explores the finite-sample properties of the inferences developed in Section 3, Section 5 proposes an application, and

Section 6 concludes. All proofs are relegated to the Appendix. Due to space considerations, additional Monte Carlo simulations and further details on some of the proofs are provided as an online supplementary appendix.

2. THE MODEL AND HYPOTHESES

We will initially be interested in developing the limiting distribution theory for a Wald-type test statistic designed to test the null hypothesis of a linear relationship between y_{t+1} and x_t against the following threshold alternative

$$y_{t+1} = \begin{cases} \alpha_1 + \beta_1 x_t + u_{t+1} & q_t \leq \gamma \\ \alpha_2 + \beta_2 x_t + u_{t+1} & q_t > \gamma, \end{cases} \tag{1}$$

where x_t is parameterized as the nearly nonstationary process

$$x_t = \rho_T x_{t-1} + v_t, \qquad \rho_T = 1 - \frac{c}{T}, \tag{2}$$

with $c > 0$, $q_t = \mu_q + u_{qt}$, and u_t, u_{qt}, and v_t are stationary random disturbances. The above parameterization allows x_t to display local to unit root behavior and has become the norm for modeling highly persistent series for which a pure unit root assumption may not always be sensible. The threshold variable q_t is taken to be a stationary process, and γ refers to the unknown threshold parameter. Under $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$, our model in (1)–(2) coincides with that in Jansson and Moreira (2006) or Campbell and Yogo (2006) and is commonly referred to as a predictive regression model, while under $\alpha_1 = \alpha_2$, $\beta_1 = \beta_2 = 0$, we have a constant mean specification.

The motivation underlying our specification in (1)–(2) is its ability to capture phenomena such as regime-specific predictability within a simple and intuitive framework. We have in mind scenarios whereby the slope corresponding to the predictor variable becomes significant solely in one regime. Alternatively, the strength of predictability may differ depending on the regime determined by the magnitude of q_t. The predictive instability in stock returns that has been extensively documented in the recent literature and the vanishing impact of DYs from the 1990s onward in particular (see Ang and Bekaert 2007 and also Table 7) may well be the consequence of the presence of regimes for instance. Among the important advantages of a threshold-based parameterization are the rich set of dynamics it allows us to capture despite its mathematical simplicity, its estimability via a simple least squares-based approach, and the observability of the variable triggering regime switches, which may help attach a "cause" to the underlying predictability. Following Petruccelli (1992), it is also useful to recall that the piecewise linear structure can be viewed as an approximation to a much wider family of nonlinear functional forms. In this sense, although we do not argue that our chosen threshold specification mimics reality, we believe that it offers a realistic approximation to a wide range of more complicated functional forms and to regime-specific behavior in particular. It is also interesting to highlight the consequences that a behavior such as (1)–(2) may have if ignored and if predictability is assessed within a linear specification instead, say $y_t = \beta x_{t-1} + u_t$. Imposing zero intercepts for simplicity and assuming (1)–(2) holds with some γ_0, it is easy to establish that $\hat{\beta} \xrightarrow{p} \beta_1 + (\beta_2 - \beta_1)P(q_t > \gamma_0)$. This raises the possibility that $\hat{\beta}$ may converge to a quantity that is very close to zero (e.g., when $P(q_t > \gamma_0) \approx \beta_1/(\beta_1 - \beta_2)$) so that tests conducted within a linear specification may frequently and wrongly suggest the absence of any predictability.

Our choice of modeling x_t as a nearly integrated process follows the same motivation as in the linear predictive regression literature, where such a choice for x_t has been advocated as an alternative to proceeding with conventional Gaussian critical values that typically provide poor finite-sample approximations to the distribution of t statistics. In the context of a stationary AR(1) for instance, Chan (1988) demonstrated that for values of $T(1-\rho) \geq 50$, the normal distribution offers a good approximation, while for $T(1-\rho) \leq 50$, the limit obtained assuming near integratedness works better when the objective involves conducting inferences about the slope parameter of the AR(1) (see also Cavanagh, Elliott, and Stock 1995 for similar points in the context of a predictive regression model). Models that combine persistent variables with nonlinear dynamics as (1)–(2) offer an interesting framework for capturing stylized facts observed in economic data. Within a univariate setting (e.g., threshold unit root models), recent contributions toward their theoretical properties have been obtained by Caner and Hansen (2001) and Pitarakis (2008).

In what follows, the threshold parameter γ is assumed unknown, with $\gamma \in \Gamma = [\gamma_1, \gamma_2]$, and γ_1 and γ_2 are selected such that $P(q_t \leq \gamma_1) = \pi_1 > 0$ and $P(q_t \leq \gamma_2) = \pi_2 < 1$, as in Caner and Hansen (2001). We also define $I_{1t} \equiv I(q_t \leq \gamma)$ and $I_{2t} \equiv I(q_t > \gamma)$ but replace the threshold variable with a uniformly distributed random variable, making use of the equality $I(q_t \leq \gamma) = I(F(q_t) \leq F(\gamma)) \equiv I(U_t \leq \lambda)$. Here, $F(.)$ is the marginal distribution of q_t, and U_t denotes a uniformly distributed random variable on $[0, 1]$. Before proceeding further, it is also useful to reformulate (1) in matrix format. Letting y denote the vector stacking y_{t+1} and X_i the matrix stacking $(I_{it} \; x_t I_{it})$ for $i = 1, 2$, we can write (1) as $y = X_1\theta_1 + X_2\theta_2 + u$ or $y = Z\theta + u$, with $Z = (X_1 \quad X_2)$, $\theta = (\theta_1, \theta_2)$, and $\theta_i = (\alpha_i, \beta_i)'$ $i = 1, 2$. For later use, we also define $X = X_1 + X_2$ as the regressor matrix that stacks the constant and x_t. It is now easy to see that for given γ or λ, the homoscedastic Wald statistic for testing a general restriction on θ, say $R\theta = 0$, is given by $W_T(\lambda) = \hat{\theta}' R'(R(Z'Z)^{-1}R')^{-1}R\hat{\theta}/\hat{\sigma}_u^2$, with $\hat{\theta} = (Z'Z)^{-1}Z'y$, and $\hat{\sigma}_u^2 = (y'y - \sum_{i=1}^{2} y'X_i(X_i'X_i)^{-1}X_i'y)/T$ is the residual variance obtained from (1). In practice, since the threshold parameter is unidentified under the null hypothesis, inferences are conducted using the SupWald formulation, expressed as $\sup_{\lambda \in [\pi_1, \pi_2]} W_T(\lambda)$, with $\pi_1 = F(\gamma_1)$ and $\pi_2 = F(\gamma_2)$. Throughout this article, the practical implementation of our SupWald statistics will use 10% trimming at each end of the sample.

In the context of our specification in (1)–(2), we will initially be interested in the null hypothesis of linearity given by H_0^A : $\alpha_1 = \alpha_2$, $\beta_1 = \beta_2$. We write the corresponding restriction matrix as $R_A = [I \quad -I]$, with I denoting a 2×2 identity matrix and the SupWald statistic $\sup_\lambda W_T^A(\lambda)$. At this stage, it is important to note that the null hypothesis given by H_0^A corresponds to the linear specification $y_{t+1} = \alpha + \beta x_t + u_{t+1}$ and thus does not test predictability per se since x_t may appear as a predictor under both the null and the alternative hypothesis. Thus, we also consider the null given by $H_0^B : \alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$, with the corresponding SupWald statistic written as $\sup_\lambda W_T^B(\lambda)$, where now $R_B = [1 \;\; 0 \;\; -1 \;\; 0, 0 \;\; 1 \;\; 0 \;\; 0, 0 \;\; 0 \;\; 0 \;\; 1]$. Under this null

hypothesis, the model is given by $y_{t+1} = \alpha + u_{t+1}$ and the test is expected to have power against departures from both linearity and predictability.

3. LARGE-SAMPLE INFERENCE

Our objective here is to investigate the asymptotic properties of Wald-type tests for detecting the presence of threshold effects in our predictive regression setup. We initially obtain the limiting distribution of $W_T^A(\lambda)$ under the null hypothesis $H_0^A : \alpha_1 = \alpha_2, \beta_1 = \beta_2$. We subsequently turn to the joint null hypothesis of linearity and no predictability given by $H_0^B : \alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$, and explore the limiting behavior of $W_T^B(\lambda)$.

Our operating assumptions about the core probabilistic structure of (1)–(2) will closely mimic the assumptions imposed in the linear predictive regression literature but will occasionally also allow for a greater degree of generality (e.g., Cavanagh et al. 1995; Campbell and Yogo 2006; Jansson and Moreira 2006, among others). Specifically, the innovations v_t will be assumed to follow a general linear process we write as $v_t = \Psi(L)e_t$, where $\Psi(L) = \sum_{j=0}^{\infty} \psi_j L^j$, $\sum_{j=0}^{\infty} j|\psi_j| < \infty$ and $\Psi(1) \neq 0$, while the shocks to y_t, denoted u_t, will take the form of a martingale difference sequence (m.d.s.) with respect to an appropriately defined information set. More formally, letting $\widetilde{w}_t = (u_t, e_t)'$ and $\mathcal{F}_t^{\widetilde{w}q} = \{\widetilde{w}_s, u_{qs} | s \leq t\}$, the filtration generated by $(\widetilde{w}_t, u_{qt})$ we will operate under the following assumptions:

Assumption 1. $E[\widetilde{w}_t | \mathcal{F}_{t-1}^{\widetilde{w}q}] = 0$, $E[\widetilde{w}_t \widetilde{w}_t' | \mathcal{F}_{t-1}^{\widetilde{w}q}] = \widetilde{\Sigma} > 0$, $\sup_t E\widetilde{w}_{it}^4 < \infty$.

Assumption 2. The threshold variable $q_t = \mu_q + u_{qt}$ has a continuous and strictly increasing distribution $F(.)$ and is such that u_{qt} is a strictly stationary, ergodic, and strong mixing sequence, with mixing numbers α_m satisfying $\sum_{m=1}^{\infty} \alpha^{\frac{1}{m} - \frac{1}{r}} < \infty$ for some $r > 2$.

One implication of Assumption 1 and the properties of $\Psi(L)$ is that a functional central limit theorem holds for the joint process $w_t = (u_t, v_t)'$ (see Phillips 1987). More formally, $\sum_{t=1}^{[Tr]} w_t / \sqrt{T} \Rightarrow B(r) = (B_u(r), B_v(r))'$, with the long-run variance of the bivariate Brownian motion $B(r)$ being given by $\Omega = \sum_{k=-\infty}^{\infty} E[w_0 w_k'] = [(\omega_u^2, \omega_{uv}), (\omega_{vu}, \omega_v^2)] = \Sigma + \Lambda + \Lambda'$. Our notation is such that $\widetilde{\Sigma} = [(\sigma_u^2, \sigma_{ue}), (\sigma_{ue}, \sigma_e^2)]$ and $\Sigma = [(\sigma_u^2, \sigma_{uv}), (\sigma_{uv}, \sigma_v^2)]$, with $\sigma_v^2 = \sigma_e^2 \sum_{j=0}^{\infty} \psi_j^2$ and $\sigma_{uv} = \sigma_{ue}$, since $E[u_t e_{t-j}] = 0 \ \forall j \geq 1$ by assumption. Given our parameterization of v_t and the m.d.s. assumption for u_t, we have $\omega_{uv} = \sigma_{ue}\Psi(1)$ and $\omega_v^2 = \sigma_e^2 \Psi(1)^2$. For later use, we also let $\lambda_{vv} = \sum_{k=1}^{\infty} E[v_t v_{t-k}]$ denote the one-sided autocovariance so that $\omega_v^2 = \sigma_v^2 + 2\lambda_{vv} \equiv \sigma_e^2 \sum_{j=0}^{\infty} \psi_j^2 + 2\lambda_{vv}$. At this stage, it is useful to note that the m.d.s. assumption in Assumption 1 imposes a particular structure on Ω. For instance, since serial correlation in u_t is ruled out, we have $\omega_u^2 = \sigma_u^2$. It is worth emphasizing, however, that while ruling out serial correlation in u_t, our assumptions allow for a sufficiently general covariance structure linking (1)–(2) and a general dependence structure for the disturbance terms driving x_t and q_t. The m.d.s. assumption on u_t is a standard assumption that has been made throughout all recent research on predictive regression models (for instance, see Campbell and Yogo 2005 and references therein; Jansson and Moreira 2006) and appears to be an intuitive operating framework, given that many applications take y_{t+1} to be stock returns. Writing $\Lambda = \sum_{k=1}^{\infty} E[w_t w_{t-k}'] = [(\lambda_{uu}, \lambda_{uv}), (\lambda_{vu}, \lambda_{vv})]$, it is also useful to explicitly highlight the fact that within our probabilistic environment, $\lambda_{uu} = 0$ and $\lambda_{uv} = 0$ due to the m.d.s. property of the $u_t's$, while λ_{vv} and λ_{vu} may be nonzero.

Regarding the dynamics of the threshold variable q_t and how it interacts with the remaining variables driving the system, Assumption 1 requires q_{t-j}'s to be orthogonal to u_t for $j \geq 1$. Since q_t is stationary, this is in a way a standard regression model assumption and is crucial for the development of our asymptotic theory. We note, however, that our assumptions allow for a broad level of dependence between the threshold variable q_t and the other variables included in the model (e.g., q_t may be contemporaneously correlated with both u_t and v_t). At this stage, it is perhaps also useful to reiterate the fact that our assumption about the correlation of q_t with the remaining components of the system is less restrictive than what is typically found in the literature on marked empirical processes or functional coefficient models such as $y_{t+1} = f(q_t)x_t + u_{t+1}$, which commonly take q_t to be independent of u_t and x_t.

Since our assumptions also satisfy Caner and Hansen's (2001) framework, from their theorem 1, we can write $\sum_{t=1}^{[Tr]} u_t I_{1t-1} / \sqrt{T} \Rightarrow B_u(r, \lambda)$ as $T \to \infty$, with $B_u(r, \lambda)$ denoting a two-parameter Brownian motion with covariance $\sigma_u^2 (r_1 \wedge r_2)(\lambda_1 \wedge \lambda_2)$ for $(r_1, r_2), (\lambda_1, \lambda_2) \in [0, 1]^2$ and where $a \wedge b \equiv \min\{a, b\}$. Noting that $B_u(r, 1) \equiv B_u(r)$, we will also make use of a particular process, known as a Kiefer process and defined as $G_u(r, \lambda) = B_u(r, \lambda) - \lambda B_u(r, 1)$. A Kiefer process on $[0, 1]^2$ is Gaussian with zero mean and covariance function $\sigma_u^2 (r_1 \wedge r_2)(\lambda_1 \wedge \lambda_2 - \lambda_1 \lambda_2)$. Finally, we introduce the diffusion process $K_c(r) = \int_0^r e^{(r-s)c} dB_v(s)$, with $K_c(r)$ such that $dK_c(r) = cK_c(r) + dB_v(r)$ and $K_c(0) = 0$. Note that we can also write $K_c(r) = B_v(r) + c \int_0^r e^{(r-s)c} B_v(s) ds$. Under our assumptions, it follows directly from lemma 3.1 in Phillips (1988) that $x_{[Tr]} / \sqrt{T} \Rightarrow K_c(r)$. For notational clarity in what follows, it is important to recall that $K_c(r)$ and all our other processes indexed by either u or v are univariate.

3.1 Testing $H_0^A : \alpha_1 = \alpha_2, \beta_1 = \beta_2$

Having outlined our key operating assumptions, we now turn to the limiting behavior of our test statistics. We will initially concentrate on the null hypothesis given by $H_0^A : \alpha_1 = \alpha_2, \beta_1 = \beta_2$ and the behavior of $\sup_\lambda W_T^A(\lambda)$, which is summarized in the following proposition.

Proposition 1. Under the null hypothesis $H_0^A : \alpha_1 = \alpha_2, \beta_1 = \beta_2$, Assumptions 1 and 2, and as $T \to \infty$, the limiting distribution of the SupWald statistic is given by:

$$\sup_\lambda W_T^A(\lambda) \Rightarrow \sup_\lambda \frac{1}{\lambda(1-\lambda)\sigma_u^2} \left[\int_0^1 \overline{K}_c(r) dG_u(r, \lambda) \right]' \times \left[\int_0^1 \overline{K}_c(r) \overline{K}_c(r)' \right]^{-1} \left[\int_0^1 \overline{K}_c(r) dG_u(r, \lambda) \right], \quad (3)$$

where $\overline{K}_c(r) = (1, K_c(r))'$, $G_u(r, \lambda)$ is a Kiefer process and $K_c(r)$ an Ornstein-Uhlenbeck process.

Although the limiting random variable in (3) appears to depend on unknown parameters, such as the correlation between B_u and B_v, σ_u^2, and the near-integration parameter c, a closer analysis of the expression suggests instead that it is equivalent to a random variable given by a quadratic form in normalized Brownian bridges, identical to the one that occurs when testing for structural breaks in a purely stationary framework. We can write it as

$$\sup_{\lambda} \frac{BB(\lambda)' BB(\lambda)}{\lambda(1-\lambda)}, \tag{4}$$

with $BB(\lambda)$ denoting a standard bivariate Brownian bridge (recall that a Brownian bridge is a zero mean Gaussian process with covariance $\lambda_1 \wedge \lambda_2 - \lambda_1\lambda_2$). This result follows from the fact that the processes $K_c(r)$ and $G_u(r, \lambda)$ appearing in the stochastic integrals in (3) are uncorrelated and thus independent since Gaussian. Indeed,

$$\begin{aligned}
&E[G_u(r_1, \lambda_1)K_c(r_2)] \\
&= E\Bigg[(B_u(r_1, \lambda_1) - \lambda_1 B_u(r_1, 1))(B_v(r_2) \\
&\quad + c\int_0^{r_2} e^{(r_2-s)c} B_v(s)ds)\Bigg] \\
&= E[B_u(r_1, \lambda_1)B_v(r_2)] - \lambda_1 E[B_u(r_1, 1)B_v(r_2)] \\
&\quad + c\int_0^{r_2} e^{(r_2-s)c} E[B_u(r_1, \lambda_1)B_v(s)]ds \\
&\quad - \lambda_1 c\int_0^{r_2} e^{(r_2-s)c} E[B_u(r_1, 1)B_v(s)]ds \\
&= \omega_{uv}(r_1 \wedge r_2)\lambda_1 - \lambda_1\omega_{uv}(r_1 \wedge r_2) \\
&\quad + c\lambda_1\int_0^{r_2} e^{(r_2-s)c}(r_1 \wedge s)ds - c\lambda_1 \\
&\quad \times \int_0^{r_2} e^{(r_2-s)c}(r_1 \wedge s)ds = 0.
\end{aligned}$$

Given that $K_c(r)$ is Gaussian and independent of $G_u(r, \lambda)$ and also $E[G_u(r_1, \lambda_1)G_u(r_2, \lambda_2)] = \sigma_u^2(r_1 \wedge r_2)((\lambda_1 \wedge \lambda_2) - \lambda_1\lambda_2$, we have $\int K_c(r)dG_u(r, \lambda) \equiv N(0, \sigma_u^2\lambda(1-\lambda)\int K_c(r)^2)$ conditionally on a realization of $K_c(r)$. Normalizing by $\sigma_u^2 \int K_c^2(r)$, as in (3), gives the Brownian bridge process in (4), which is also the unconditional distribution since it is not dependent on a realization of $K_c(r)$ (see also lemma 5.1 in Park and Phillips 1988). Obviously, the discussion trivially carries through to $\overline{K}_c$ and G_u since $E[\overline{K}_c(r_2)G_u(r_1, \lambda_1)]' = E[G_u(r_1, \lambda_1) \;\; K_c(r_2)G_u(r_1, \lambda_1)]' = [0 \;\; 0]'$.

The result in Proposition 1 is unusual and interesting for a variety of reasons. It highlights an environment in which the null distribution of the SupWald statistic no longer depends on any nuisance parameters as it is typically the case in a purely stationary environment and thus no bootstrapping schemes are needed for conducting inferences. In fact, the distribution presented in Proposition 1 is extensively tabulated in Andrews (1993), and Hansen (1997) provided p-value approximations that can be used for inference purposes. More recently, Estrella (2003) also provided exact p-values for the same distribution. Finally and perhaps more importantly, the limiting distribution does not depend on c, the near-integration parameter, which is another unusual feature of our framework.

All these properties are in contrast with what has been documented in the recent literature on testing for threshold effects in purely stationary contexts. In Hansen (1996) for instance, the author investigated the limiting behavior of a SupLM-type test statistic for detecting the presence of threshold nonlinearities in purely stationary models. There, it was established that the key limiting random variables depend on numerous nuisance parameters involving unknown population moments of variables included in the fitted model. From theorem 1 in Hansen (1996), it is straightforward to establish, for instance, that under stationarity, the limiting distribution of a Wald-type test statistic would be given by $S^*(\lambda)'M^*(\lambda)^{-1}S^*(\lambda)$, with $M^*(\lambda) = M(\lambda) - M(\lambda)M(1)^{-1}M(\lambda)$, and $S^*(\lambda) = S(\lambda) - M(\lambda)M(1)^{-1}S(1)$. Here, $M(\lambda) = E[X_1'X_1]$ and $S(\lambda)$ is a zero mean Gaussian process with variance $M(\lambda)$. Since in this context the limiting distribution depends on the unknown model-specific population moments, the practical implementation of inferences is through a bootstrap-style methodology.

One interesting instance worth pointing out, however, is the fact that this limiting random variable simplifies to a Brownian bridge type of limit when the threshold variable is taken as exogenous in the sense $M(\lambda) = \lambda M(1)$. Although the comparison with the present context is not obvious since we take x_t to be near-integrated and we allow the innovations in q_t to be correlated with those of x_t, the force behind the analogy comes from the fact that x_t and q_t have variances with different orders of magnitude. In a purely stationary setup, taking x_t as stationary and the threshold variable as some uniformly distributed random variable leads to results such as $\sum x_t^2 I(U_t \leq \lambda)/T \xrightarrow{p} E[x_t^2 I(U_t \leq \lambda)]$, and if x_t and U_t are independent, we also have $E[x_t^2 I(U_t \leq \lambda)] = \lambda E[x_t^2]$. It is this last key simplification that is instrumental in leading to the Brownian bridge type of limit in Hansen's (1996) framework. If now x_t is taken as a nearly integrated process and regardless of whether its shocks are correlated with U_t or not, we have $\sum x_t^2 I(U_t \leq \lambda)/T^2 \Rightarrow \lambda \int K_c^2(r)$, which can informally be viewed as analogous to the previous scenario. Heuristically, this result follows by establishing that asymptotically, objects interacting $x_t/\sqrt{T}$ and $(I_{1t} - \lambda)$, such as $\frac{1}{T}\sum(\frac{x_t}{\sqrt{T}})^2(I_{1t} - \lambda)$ or $\frac{1}{T}\sum(\frac{x_t}{\sqrt{T}})(I_{1t} - \lambda)$, converge to zero (see also Caner and Hansen 2001, p. 1585; Pitarakis 2008). This would be similar to arguing that $x_t/\sqrt{T}$ and I_{1t} are asymptotically uncorrelated in the sense that their sample covariance (normalized by T) is zero in the limit.

3.2 Testing $H_0^B : \alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$

We next turn to the case where the null hypothesis of interest tests jointly the absence of linearity and no predictive power; that is, we focus on testing $H_0^B : \alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$ using the supremum of $W_T^B(\lambda)$. The following proposition summarizes its limiting behavior.

Proposition 2. Under the null hypothesis $H_0^B : \alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$, Assumptions 1 and 2, and as $T \to \infty$, the

limiting distribution of the SupWald statistic is given by

$$\sup_{\lambda} W_T^B(\lambda) \Rightarrow \frac{\left[\int K_c^*(r) dB_u(r,1)\right]^2}{\sigma_u^2 \int K_c^*(r)^2} + \sup_{\lambda} \frac{1}{\lambda(1-\lambda)\sigma_u^2} \times \left[\int \overline{K}_c^*(r) dG_u(r,\lambda)\right]' \left[\int \overline{K}_c^* \overline{K}_c^*(r)'\right]^{-1} \times \left[\int \overline{K}_c^*(r) dG_u(r,\lambda)\right]', \quad (5)$$

where $\overline{K}_c^*(r) = (1, K_c^*(r))'$, $K_c^*(r) = K_c(r) - \int_0^1 K_c(r)dr$ and the remaining variables are as in Proposition 1.

Looking at the expression of the limiting random variable in (5), we note that it consists of two components, with the second one being equivalent to the limiting random variable we obtained under Proposition 1. Under endogeneity, the first component on the right-hand side of (5) is more problematic in the sense that it does not simplify further due to the fact that $K_c^*(r)$ and $B_u(r,1)$ are correlated, since ω_{uv} may take nonzero values. However, if we were to rule out endogeneity by setting $\omega_{uv} = 0$, then it is interesting to note that the limiting distribution of the SupWald statistic in Proposition 2 takes the following simpler form

$$\sup_{\lambda} W_T^B(\lambda) \Rightarrow W(1)^2 + \sup_{\lambda} \frac{BB(\lambda)' BB(\lambda)}{\lambda(1-\lambda)}, \quad (6)$$

where $BB(\lambda)$ is a bivariate Brownian bridge and $W(1)$ a univariate standard normally distributed random variable. The first component on the right-hand side of either (5) or (6) can be recognized as the $\chi^2(1)$ limiting distribution of the Wald statistic for testing $H_0 : \beta = 0$ in the linear specification

$$y_{t+1} = \alpha + \beta x_t + u_{t+1}, \quad (7)$$

and the presence of this first component makes the test powerful in detecting deviations from the null (see Rossi 2005 for an illustration of a similar phenomenon in a different context).

Our next concern is to explore ways of making (5) operational since, as it stands, the first component of the limiting random variable depends on model-specific moments and cannot be universally tabulated. For this purpose, it is useful to note that the problems arising from the practical implementation of (5) are partly analogous to the difficulties documented in the single-equation cointegration testing literature, where the goal was to obtain nuisance-parameter-free chi-squared asymptotics for Wald-type tests on β in (7) despite the presence of endogeneity (see Phillips and Hansen 1990; Saikkonen 1991, 1992). As shown by Elliott (1998), however, inferences about β in (7) can no longer be mixed normal when x_t is a near-unit root process. It is only very recently that Phillips and Magdalinos (2009) (PM09 thereafter) reconsidered the issue and resolved the difficulties discussed in Elliott (1998) via the introduction of a new instrumental-variable-type estimator of β in (7). Their method is referred to as IVX estimation since the relevant IV is constructed solely via a transformation of the existing regressor x_t. It is this same method that we propose to adapt to our present context.

Before proceeding further, it is useful to note that $W_T^B(\lambda)$ can be expressed as the sum of the following two components:

$$W_T^B(\lambda) \equiv \frac{\hat{\sigma}_{\text{lin}}^2}{\hat{\sigma}_u^2} W_T(\beta = 0) + W_T^A(\lambda), \quad (8)$$

where $W_T(\beta = 0)$ is the standard Wald statistic for testing $H_0 : \beta = 0$ in (7) and $W_T^A(\lambda)$ is as in Proposition 1. Specifically,

$$W_T(\beta = 0) = \frac{1}{\hat{\sigma}_{\text{lin}}^2} \frac{\left[\sum x_{t-1} y_t - T\bar{x}\bar{y}\right]^2}{\left[\sum x_{t-1}^2 - T\bar{x}^2\right]}, \quad (9)$$

with $\bar{x} = \sum x_{t-1}/T$ and $\hat{\sigma}_{\text{lin}}^2 = (y'y - y'X(X'X)^{-1}X'y)/T$ is the residual variance obtained from the same linear specification. Although not of direct interest, this reformulation of $W_T^B(\lambda)$ can simplify the implementation of the IVX version of the Wald statistic since the setup is now identical to that of PM09 and involves constructing a Wald statistic for testing $H_0 : \beta = 0$ in (7); that is, we replace $W_T(\beta = 0)$ in (8) with its IVX-based version, which is shown to be asymptotically distributed as a $\chi^2(1)$ random variable that does not depend on the noncentrality parameter c or other endogeneity-induced parameters. Note that although PM09 operated within a model without an intercept, Stamatogiannis (2010) and Kostakis, Magdalinos, and Stamatogiannis (2010) have also established the validity of the theory in models with a fitted constant term.

The IVX methodology starts by choosing an artificial slope coefficient, say

$$R_T = 1 - \frac{c_z}{T^\delta} \quad (10)$$

for a given constant $c_z > 0$ and $\delta < 1$ and uses the latter to construct an IV generated as $\tilde{z}_t = R_T \tilde{z}_{t-1} + \Delta x_t$ or under zero initialization $\tilde{z}_t = \sum_{j=1}^t R_T^{t-j} \Delta x_j$. This IV is then used to obtain an IV estimator of β in (7) and to construct the corresponding Wald statistic for testing $H_0 : \beta = 0$. Through this judicious choice of the instrument, PM09 showed that it is possible to clean out the effects of endogeneity even within the near-unit root case and to subsequently obtain an estimator of β that is mixed normal under a suitable choice of δ (i.e., $\delta \in (2/3, 1)$) and setting $c_z = 1$ (see PM09, pp. 7–12). More importantly, the resulting limiting distribution of the Wald statistic for testing $\beta = 0$ in (7) no longer depends on the noncentrality parameter c.

Following PM09 and Stamatogiannis (2010) and letting y_t^*, x_t^*, and $\tilde{z}_t^*$ denote the demeaned versions of y_t, x_t, and $\tilde{z}_t$, we can write the IVX estimator of β as $\tilde{\beta}^{\text{ivx}} = \sum y_t^* \tilde{z}_{t-1}^* / \sum x_{t-1}^* \tilde{z}_{t-1}^*$. Note that contrary to PM09 or Stamatogiannis (2010), we do not need a bias correction term in the numerator of $\tilde{\beta}^{\text{ivx}}$ since we operate under the assumption that $\lambda_{uv} = 0$. The corresponding IVX-based Wald statistic for testing $H_0 : \beta = 0$ in (7) is now written as:

$$W_T^{\text{ivx}}(\beta = 0) = \frac{(\tilde{\beta}^{\text{ivx}})^2 \left(\sum x_{t-1}^* \tilde{z}_{t-1}^*\right)^2}{\tilde{\sigma}_u^2 \sum \left(\tilde{z}_{t-1}^*\right)^2}, \quad (11)$$

with $\tilde{\sigma}_u^2 = \sum (y_t^* - \tilde{\beta}^{\text{IVX}} x_{t-1}^*)^2 / T$. Note that this latter quantity is also asymptotically equivalent to $\hat{\sigma}_{\text{lin}}^2$ since the least squares estimator of β remains consistent. Under the null hypothesis H_0^B, we also have that these two residual variances are in turn asymptotically equal to $\hat{\sigma}_u^2$ so that $\hat{\sigma}_{\text{lin}}^2 / \hat{\sigma}_u^2 \approx 1$ in (8).

We can now introduce our modified Wald statistic, say $W_T^{B,\text{ivx}}(\lambda)$, for testing $H_0^B : \alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$ in (1) as

$$W_T^{B,\text{ivx}}(\lambda) = W_T^{\text{ivx}}(\beta = 0) + W_T^A(\lambda). \tag{12}$$

Its limiting behavior is summarized in the following proposition.

Proposition 3. Under the null hypothesis $H_0^{(B)} : \alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$, Assumptions 1 and 2, $\delta \in (2/3, 1)$ in (10), and as $T \to \infty$, we have

$$\sup_\lambda W_T^{B,\text{ivx}}(\lambda) \Rightarrow W(1)^2 + \sup_\lambda \frac{BB(\lambda)'BB(\lambda)}{\lambda(1-\lambda)}, \tag{13}$$

with $BB(\lambda)$ denoting a standard Brownian bridge.

Our result in (13) highlights the usefulness of the IVX-based estimation methodology since the resulting limiting distribution of the SupWald statistic is now equivalent to the one obtained under strict exogeneity (i.e., under $\omega_{uv} = 0$) in (6). The practical implementation of the test is also straightforward, requiring nothing more than the computation of an IV estimator.

4. FINITE-SAMPLE ANALYSIS

4.1 Testing $H_0^A : \alpha_1 = \alpha_2, \beta_1 = \beta_2$

Having established the limiting properties of the SupWald statistic for testing H_0^A, our next goal is to illustrate the finite-sample adequacy of our asymptotic approximation and empirically illustrate our theoretical findings. It will also be important to highlight the equivalence of the limiting results obtained in Proposition 1 to the Brownian bridge type of limit documented in Andrews (1993) and for which Hansen (1997) obtained p-value approximations and Estrella (2003), exact p-values. Naturally, this allows us to evaluate the size properties of our tests as well.

Our data-generating process (DGP) under H_0^A is given by the following set of equations:

$$\begin{aligned} y_t &= \alpha + \beta x_{t-1} + u_t, \\ x_t &= \left(1 - \frac{c}{T}\right) x_{t-1} + v_t, \\ v_t &= \rho v_{t-1} + e_t, \end{aligned} \tag{14}$$

with u_t and e_t both NID(0, 1), while the fitted model is given by (1), with q_t assumed to follow the AR(1) process $q_t = \phi q_{t-1} + u_{qt}$, with $u_{qt} =$ NID(0, 1). Regarding the covariance structure of the random disturbances, letting $z_t = (u_t, e_t, u_{qt})'$ and $\mathbf{\Sigma}_z = E[z_t z_t']$, we use

$$\mathbf{\Sigma}_z = \begin{pmatrix} 1 & \sigma_{ue} & \sigma_{uu_q} \\ \sigma_{ue} & 1 & \sigma_{eu_q} \\ \sigma_{uu_q} & \sigma_{eu_q} & 1 \end{pmatrix},$$

which allows for a sufficiently general covariance structure, while imposing unit variances. Note also that our chosen covariance matrix parameterization allows the threshold variable to be contemporaneously correlated with the shocks to y_t. All our H_0^A-based size experiments use $N = 5000$ replications and set $\{\alpha, \beta, \rho, \phi\} = \{0.01, 0.10, 0.40, 0.50\}$ throughout. Since our initial motivation is to explore the theoretically documented robustness of the limiting distribution of SupWald^A to the presence or absence of endogeneity, we consider the two scenarios given by

$$\text{DGP}_1 : \{\sigma_{ue}, \sigma_{uu_q}, \sigma_{eu_q}\} = \{-0.5, 0.3, 0.4\},$$
$$\text{DGP}_2 : \{\sigma_{ue}, \sigma_{uu_q}, \sigma_{eu_q}\} = \{0.0, 0.0, 0.0\}.$$

The implementation of all our Sup-based tests assumes 10% trimming at each end of the sample.

Table 1 presents some key quantiles of the SupWald^A distribution (see Proposition 1) simulated using moderately small sample sizes and compares them with their asymptotic counterparts. Results are displayed solely for the DGP_1 covariance structure since the corresponding figures for DGP_2 were almost identical.

Looking across the different values of c as well as the different quantiles, we note an excellent adequacy of the $T = 200$- and $T = 400$-based finite-sample distributions to the asymptotic counterpart tabulated in Andrews (1993) and Estrella (2003). This also confirms our results in Proposition 1 and provides empirical support for the fact that inferences are robust to the magnitude of c. Note that with $T = 200$, the values of $(1 - c/T)$ corresponding to our choices of c in Table 1 are 0.995, 0.975, 0.950, and 0.800, respectively. Thus, the quantiles of the simulated distribution appear to be highly robust to a wide range of persistence characteristics.

Naturally, the fact that our finite-sample quantiles match closely with their asymptotic counterparts even under $T = 200$ is not sufficient to claim that the test has good size properties. For this purpose, we have computed the empirical size of the SupWald^A-based test, making use of the pvsup routine of Hansen (1997). The latter is designed to provide approximate p-values for test statistics whose limiting distribution is as in (4). Results are presented in Table 2, which concentrates solely on the DGP_1 covariance structure. We initially focus on the first two

Table 1. Critical values of SupWald^A

	DGP_1, T=200				DGP_1, T=400				
	$c = 1$	$c = 5$	$c = 10$	$c = 20$	$c = 1$	$c = 5$	$c = 10$	$c = 20$	∞
2.5%	2.18	2.21	2.21	2.19	2.31	2.24	2.24	2.27	2.41
5.0%	2.53	2.52	2.57	2.50	2.65	2.63	2.62	2.63	2.75
10.0%	3.01	3.07	2.99	2.99	3.13	3.10	3.11	3.12	3.27
90.0%	10.20	10.46	10.48	10.39	10.28	10.23	10.20	10.30	10.46
95.0%	12.07	12.03	12.13	12.19	11.85	12.05	12.11	12.08	12.17
97.5%	13.82	13.76	13.85	13.84	13.74	13.57	13.91	13.64	13.71

Table 2. Size properties of SupWald^A

	$T = 200$			$T = 400$			$T = 200$, BOOT			$T = 400$, BOOT		
	2.5%	5.0%	10%	2.5%	5.0%	10%	2.5%	5.0%	10.0%	2.5%	5.0%	10.0%
$c = 1$	2.60	4.70	8.90	2.50	4.60	9.60	3.01	6.20	11.14	3.62	5.98	11.02
$c = 5$	2.50	4.90	9.30	2.40	4.90	9.30	2.98	6.36	11.86	3.38	6.08	11.02
$c = 10$	2.80	4.80	9.20	2.70	5.10	9.30	3.26	6.42	12.00	3.26	5.64	10.66
$c = 20$	2.60	4.80	9.50	2.50	5.00	9.60	3.20	6.42	11.32	3.26	6.16	11.40

left-hand panels, while the ones referred to as $T = 200$, BOOT and $T = 400$, BOOT are discussed later.

From the figures presented in the two left-hand panels in Table 2, we again note the robustness of the empirical size estimates of SupWald^A to the magnitude of the noncentrality parameter. Overall, the size estimates match their nominal counterparts quite accurately even under a moderately small sample size.

It is also interesting to compare the asymptotic approximation in (4) with that occurring when x_t is assumed to follow an AR(1), with $|\rho| < 1$, rather than the local to unit root specification we have adopted in this article. Naturally, under pure stationarity, the results of Hansen (1996, 1999) apply and inferences can be conducted by simulating critical values from the asymptotic distribution that is the counterpart of (3) obtained under pure stationarity and following the approach outlined in the aforementioned articles. This latter approach is similar to an external bootstrap but should not be confused with the idea of obtaining critical values from a bootstrap distribution. The obvious question we are next interested in documenting is which approximation works better when x_t is a highly persistent process. For this purpose, the two right-hand panels in Table 2, referred to as BOOT, present the corresponding empirical size estimates obtained using the asymptotic approximation and its external bootstrap-style implementation developed by Hansen (1996, 1999) and justified by the multiplier central limit theorem (see Van der Vaart and Wellner 1996). Although our comparison involves solely size properties, the above figures suggest that our nuisance-parameter-free Brownian bridge-based asymptotic approximation does a good job in matching empirical sizes with nominal sizes when ρ is close to the unit root frontier. Proceeding using Hansen's (1996) approach on the other hand suggests that the procedure is mildly oversized, which does not taper off as T is allowed to increase.

Before proceeding further, it is also important to document SupWald^A's ability to correctly detect the presence of threshold effects via a finite-sample power analysis. Our goal here is not to develop a full theoretical and empirical power analysis of our test statistics that would bring us well beyond our scope but to instead give a snapshot of the ability of our test statistics to lead to a correct decision under a series of fixed departures from the null. All our power-based DGPs use the same covariance structure as our size experiments and are based on the following configurations for $\{\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma\}$ in (1): DGP$_1^A$ $\{-0.03, -0.03, 1.26, 1.20, 0\}$, DGP$_2^A$ $\{-0.03, 0.15, 1.26, 1.20, 0\}$, and DGP$_3^A$ $\{-0.03, 0.25, 1.26, 1.26, 0\}$, thus covering both intercept only, slope only, and joint intercept and slope shifts. In Table 3, the figures represent correct decision frequencies evaluated as the number of times the p-value of the test statistic leads to a rejection of the null using a 2.5% nominal level.

We note from Table 3 that power converges toward 1 under all three parameter configurations, albeit quite slowly when only intercepts are characterized by threshold effects. The test displays good finite-sample power even under $T = 200$ when the slopes are allowed to shift, as in DGP$_1^A$ and DGP$_2^A$. It is also interesting to note the negative influence of an increasing c on finite-sample power under the DGPs with shifting slopes. As expected, this effect vanishes asymptotically since even for $T \geq 400$, the frequencies across the different magnitudes of c become very similar.

4.2 Testing $H_0^B : \alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$

We next turn to the null hypothesis given by $H_0^B : \alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$. As documented in Proposition 2, we recall that the limiting distribution of the SupWald^B statistic is no longer free of nuisance parameters and does not take a familiar form when we operate under the set of assumptions characterizing Proposition 1. However, one instance under which the limiting distribution of the SupWald^B statistic takes a simple form is when we impose the exogeneity assumption, as when considering the covariance structure referred to as DGP$_2$ above. Under this scenario, the relevant limiting distribution is given by (6) and can be easily tabulated through standard simulation-based methods.

Table 3. Power properties of SupWald^A

	$c = 1$			$c = 5$			$c = 10$		
	DGP$_1^A$	DGP$_2^A$	DGP$_3^A$	DGP$_1^A$	DGP$_2^A$	DGP$_3^A$	DGP$_1^A$	DGP$_2^A$	DGP$_3^A$
$T = 200$	0.73	0.73	0.15	0.39	0.44	0.14	0.20	0.26	0.14
$T = 400$	0.98	0.98	0.37	0.92	0.93	0.37	0.78	0.82	0.37
$T = 1000$	1.00	1.00	0.88	1.00	1.00	0.89	1.00	1.00	0.86

Table 4. Critical values of SupWald^B under exogeneity

	2.5%	5%	10%	90%	95%	97.5%
			$c = 1$			
$T = 200$	2.59	3.03	3.58	11.73	13.63	15.36
$T = 400$	2.67	3.06	3.67	11.80	13.69	15.41
$T = 800$	2.67	3.15	3.78	11.71	13.42	15.35
			$c = 5$			
$T = 200$	2.56	3.02	3.64	11.63	13.69	15.46
$T = 400$	2.65	3.06	3.69	11.97	13.79	15.85
$T = 800$	2.71	3.15	3.73	11.55	13.42	15.14

For this purpose, Table 4 presents some empirical quantiles obtained using $T = 200$, $T = 400$, and $T = 800$ from the null DGP $y_t = 0.01 + u_t$. As can be inferred from (6) we note that the quantiles are unaffected by the chosen magnitude of c and appear sufficiently stable across the different sample sizes considered. Viewing the $T = 800$-based results as approximating the asymptotic distribution for instance, the quantiles obtained under $T = 200$ and $T = 400$ match closely with their asymptotic counterparts.

We next turn to the more general scenario in which one wishes to test H_0^B within a specification that allows for endogeneity. Taking our null DGP as $y_t = 0.01 + u_t$ and the covariance structure referred to as DGP$_1$, it is clear from Proposition 2 that using the critical values from Table 4 will lead to misleading results. This is indeed confirmed empirically with size estimates for SupWald^B lying about two percentage points above their nominal counterparts (see Table 5). Using our IVX-based test statistic in (11)–(12), however, ensures that the above critical values remain valid even under the presence of endogeneity. Results for this experiment are also presented in Table 5. Table 5 also aims to highlight the influence of the choice of the δ parameter in the construction of the IVX variable (see (10)) on the size properties of the test.

Overall, we note an excellent match of the empirical sizes with their nominal counterparts. As δ increases toward 1, it is possible to note a very slight deterioration in the size properties of SupWald$^{B,\text{ivx}}$, with empirical sizes mildly exceeding their nominal counterparts. Looking also at the power figures presented in Table 6, it is clear that as $\delta \to 1$, there is a very mild-size power tradeoff that kicks in. This is perhaps not surprising since as $\delta \to 1$, the instrumental variable starts behaving like the original nearly integrated regressor. Overall, choices of δ in the 0.7–0.8 region appear to lead to very sensible results, with almost unnoticeable variations in the corresponding size estimates. Even under $\delta = 0.9$ and looking across all configurations, we can reasonably argue that the resulting size properties are good to excellent. Finally, the rows labeled SupWald^B clearly highlight the unsuitability of this uncorrected test statistic, whose limiting distribution is as in (5) and is affected by the presence of endogeneity as well as the near-integration parameter c in the underlying model. In additional simulations not reported here for instance and a configuration given by $\{\sigma_{ue}, \sigma_{uu_q}, \sigma_{eu_q}\} = \{-0.7, 0.3, 0.3\}$, $T = 200$, $\{c, \delta\} = \{1, 0.7\}$, we obtained empirical size estimates of 4.44%, 8.28%, and 15.04% under 2.5%, 5%, and 10% nominal sizes, respectively, for SupWald^B compared with 2.78%, 5.60%, and 10.70% for SupWald$^{B,\text{ivx}}$.

Next, we also considered the finite-sample power properties of our SupWald$^{B,\text{ivx}}$ statistic through a series of fixed departures from the null based on the following configurations for $\{\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma\}$: DGP$_1^B$ $\{0.01, 0.01, 0.05, 0.05, 0\}$, DGP$_2^B$ $\{-0.03, 0.25, 0.05, 0.05, 0\}$, and DGP$_3^B$ $\{0.01, 0.25, 0, 0, 0\}$. Results for this set of experiments are presented in Table 6.

The above figures suggest that our modified SupWald$^{B,\text{ivx}}$ statistic has good power properties under moderately large sample sizes. We again note that violating the null restriction that affects the slopes leads to substantially better power properties than scenarios where solely the intercepts violate the equality constraint.

5. REGIME-SPECIFIC PREDICTABILITY OF RETURNS WITH VALUATION RATIOS

One of the most frequently explored specification in the financial economics literature has aimed to uncover the predictive

Table 5. Size properties of SupWald$^{B,\text{ivx}}$ and SupWald^B under endogeneity

	2.5%	5.0%	10.0%	2.5%	5.0%	10.0%	2.5%	5%	10%
$T = 200$		$c = 1$			$c = 5$			$c = 10$	
$\delta = 0.70$	2.80	5.12	10.26	2.48	5.02	10.40	2.62	5.00	10.34
$\delta = 0.80$	2.84	5.60	10.38	2.52	5.08	10.78	2.70	5.10	10.40
$\delta = 0.90$	3.04	5.48	10.68	2.70	5.20	10.86	2.76	5.32	10.56
SupWald^B	3.54	6.36	12.28	3.06	5.94	11.52	2.98	5.72	11.14
$T = 400$		$c = 1$			$c = 5$			$c = 10$	
$\delta = 0.70$	3.02	5.66	11.06	3.00	5.36	10.60	2.74	5.32	10.14
$\delta = 0.80$	3.14	5.92	11.46	3.14	5.36	10.94	2.82	5.44	10.32
$\delta = 0.90$	3.42	6.28	12.08	3.24	5.52	11.04	2.82	5.48	10.52
SupWald^B	4.28	7.30	13.20	3.46	6.22	11.46	3.08	5.66	11.08
$T = 1000$		$c = 1$			$c = 5$			$c = 10$	
$\delta = 0.70$	2.74	5.14	10.24	2.62	4.96	10.22	2.50	4.72	10.18
$\delta = 0.80$	2.96	5.68	10.74	2.64	5.40	10.12	2.66	4.74	10.62
$\delta = 0.90$	3.30	5.90	11.50	2.92	5.42	10.06	2.64	4.96	10.44
SupWald^B	4.00	6.52	13.18	3.22	5.72	10.74	2.74	5.16	10.74

Table 6. Power properties of SupWald$^{B,\text{ivx}}$

	DGP_1^B			DGP_2^B			DGP_3^B		
$c = 1, T$	200	400	1000	200	400	1000	200	400	1000
$\delta = 0.70$	0.81	0.97	1.00	0.89	0.99	1.00	0.17	0.37	0.87
$\delta = 0.80$	0.89	0.99	1.00	0.94	1.00	1.00	0.17	0.37	0.87
$c = 5, T$	200	400	1000	200	400	1000	200	400	1000
$\delta = 0.70$	0.71	1.00	1.00	0.85	1.00	1.00	0.16	0.36	0.87
$\delta = 0.80$	0.79	1.00	1.00	0.89	1.00	1.00	0.16	0.36	0.87
$c = 10, T$	200	400	1000	200	400	1000	200	400	1000
$\delta = 0.70$	0.51	1.00	1.00	0.74	1.00	1.00	0.16	0.36	0.87
$\delta = 0.80$	0.58	1.00	1.00	0.78	1.00	1.00	0.16	0.36	0.86

power of valuation ratios such as DYs for future stock returns via significance tests implemented on simple linear regressions linking r_{t+1} to DY_t. The econometric complications that arise due to the presence of a persistent regressor, together with endogeneity issues, have generated a vast methodological literature aiming to improve inferences in such models, commonly referred to as predictive regressions (e.g., Valkanov 2003; Lewellen 2004; Campbell and Yogo 2006; Jansson and Moreira 2006; Ang and Bekaert 2007, among numerous others).

Given the multitude of studies conducted over a variety of sample periods, methodologies, data definitions, and frequencies, it is difficult to extract a clear consensus on predictability. From the recent analysis of Campbell and Yogo (2006), there appears to be statistical support for some very mild DY-based predictability, with the latter having substantially declined in strength post 1995 (see also Lettau and Van Nieuwerburgh 2008). Using monthly data over the 1946–2000 period, Lewellen (2004) documented a rather stronger DY-based predictability using a different methodology that was mainly concerned with small-sample bias correction. See also Cochrane (2008) for a more general overview of this literature.

Our goal here is to reconsider this potential presence of predictability through our regime-based methodology focusing on the DY predictor. More specifically, using growth in industrial production as our threshold variable proxying for aggregate macro conditions, our aim is to assess whether the data support the presence of regime-dependent predictability induced by good versus bad economic times. Theoretical arguments justifying the possible existence of episodic instability in predictability have been alluded to in the theoretical setting of Menzly, Santos, and Veronesi (2004), and more recently, Henkel, Martin, and Nardari (2009) explored the issue empirically using Bayesian methods within a Markov switching setup. We will show that our approach leads to a novel view and interpretation of the predictability phenomenon and that its conclusions are robust across alternative sample periods. Moreover, our findings may provide an explanation for the lack of robustness to the sample period documented in existing work under linearity. An alternative strand of the recent predictive regression literature, or more generally the forecasting literature, has also explored the issue of predictive instability through the allowance of time variation via structural breaks and the use of recursive estimation techniques. A general message that has come out from this research is the omnipresence of model instability and the important influence of time variation on forecasts (see Rossi 2005, 2006; Rapach and Wohar 2006; Timmermann 2008, among others). Our own research is also motivated by similar concerns but focuses on explicitly identifying predictability episodes induced by a particular variable such as a business cycle proxy.

Our analysis will be based on the same CRSP (Center for Research in Security Prices) dataset as the one considered in the vast majority of predictability studies (value-weighted returns for NYSE, AMEX, and NASDAQ). Throughout all our specifications, the dividend yield is defined as the aggregate dividends paid over the last 12 months divided by the market capitalization and is logged throughout (LDY thereafter). For robustness considerations, we will distinguish between returns that include dividends and returns that exclude dividends. Finally, using the 90-day T-bills (Treasury bills), all our inferences will also distinguish between raw returns and their excess counterparts. Following Lewellen (2004), we will restrict our sample to the post-war period. We will concentrate solely on monthly data since the regime-specific nature of our models would make yearly or even quarterly data-based inferences less reliable due to the potentially very small size of the sample. We will subsequently explore the robustness of our results to alternative sample periods.

Looking first at the stochastic properties of the DY predictor over the 1950M1–2007M12 period, it is clear that the series is highly persistent, as judged by a first-order sample autocorrelation coefficient of 0.991. A unit root test implemented on the same series unequivocally fails to reject the unit root null. The industrial production growth series is stationary as expected, displaying some very mild first-order serial correlation and clearly conforming to our assumptions about q_t in (1)–(2). Before proceeding with the detection of regime-specific predictability, we start by assessing return predictability within a linear specification, as has been done in the existing literature. Results across both raw and excess returns are presented in Table 7, with VWRETD denoting the returns inclusive of dividends and VWRETX denoting the returns ex-dividends. The columns named as p and p_{HAC} refer to the standard and HAC (heteroscedasticity and autocorrelation consistent)-based p-values.

The coefficient estimates of Table 7 refer to the OLS (ordinary least squares) estimates of β_{DY} in the regression $r_{t+1} = \alpha + \beta_{\text{DY}}\ \text{LDY}_t + u_{t+1}$. Focusing first on the VWRETD series, our results conform with the consensus that predictability has been vanishing from the late 1980s onward (for instance, see Campbell and Yogo 2006). The remaining p-values suggest

Table 7. Linear predictability $r_{t+1} = \alpha_{DY} + \beta_{DY}LDY_t + u_{t+1}$

VWRETD	$\hat{\beta}_{DY}$	p_{HAC}	p	R^2	VWRETX	$\hat{\beta}_{DY}$	p_{HAC}	p	R^2
1950–2007	0.010	0.011	0.008	0.9%	1950–2007	0.008	0.054	0.046	0.4%
1960–2007	0.010	0.056	0.037	0.6%	1960–2007	0.008	0.142	0.110	0.3%
1970–2007	0.009	0.069	0.056	0.6%	1970–2007	0.007	0.170	0.148	0.2%
1980–2007	0.011	0.059	0.042	0.9%	1980–2007	0.009	0.131	0.103	0.5%
1990–2007	0.014	0.153	0.105	0.8%	1990–2007	0.001	0.207	0.152	0.5%
Excess					Excess				
1950–2007	0.009	0.025	0.019	0.7%	1950–2007	0.007	0.102	0.087	0.3%
1960–2007	0.007	0.210	0.169	0.2%	1960–2007	0.004	0.417	0.372	0.0%
1970–2007	0.006	0.269	0.240	0.1%	1970–2007	0.004	0.665	0.479	0.0%
1980–2007	0.007	0.253	0.208	0.2%	1980–2007	0.005	0.439	0.392	0.0%
1990–2007	0.013	0.198	0.138	0.6%	1990–2007	0.011	0.263	0.196	0.0%

some mild predictability, especially when considering the entire 1950–2007 sample range. Interestingly, as we switch from raw to excess returns, the picture changes considerably, with most p-values strongly pointing toward the absence of any predictability. Given these p-value magnitudes, it is difficult to conceive that any methodological improvements may reverse the big picture. Also worth pointing out is the fact that a conventional test for heteroscedasticity implemented on the above specifications failed to reject the null of no heteroscedasticity. This is particularly reassuring since one of our assumptions leading to our theoretical results in Propositions 1 and 2 ruled out the presence of heteroscedasticity.

Next, focusing on the returns that exclude dividend payments, it is again the case that with p-values as high as 0.665, the null of no predictability cannot be rejected. Results appear to also be robust across different starting periods, except perhaps under the full 1950–2007 range, under which we note a mild rejection of the null. It is also important to note that all results were robust across HAC versus non-HAC standard errors. This latter point is particularly important since our assumptions surrounding (1)–(2) rule out serial correlation and heteroscedasticity in u_t.

Overall, the above linearity-based results corroborate the view that predictability is at best mildly present and its strength appears to have declined. Perhaps more importantly, Table 7 also suggests that one should be particularly cautious and worry about robustness considerations when assessing DY-induced predictability of returns since findings may be extremely sensitive to data definitions, frequency, and chosen sample period. At this stage, it is also important to reiterate that our analysis in Table 7 is mainly meant to provide a comparison benchmark for our subsequent regime-based inferences rather than reverse findings from the existing literature. This is also the reason why we do not explore outcomes based on alternative methodologies, as developed in the recent econometric literature.

The fact that numerous studies documented a decline in predictability characterizing the 1990s could also be due to the fact that predictability kicks in during particular economic episodes. Table 8 presents the results of our tests of the hypotheses $H_0^B : \alpha_1 = \alpha_2, \beta_1 = \beta_2 = 0$ and $H_0^A : \alpha_1 = \alpha_2, \beta_1 = \beta_2$ as applied to the VWRETD series ($*$ indicates rejection at 2.5%). Since results for the return series that exclude dividends as well as their excess counterparts were both qualitatively and quantitatively similar, in what follows, we concentrate solely on the VWRETD series.

The evidence presented in Table 8 comfortably points toward the presence of regime-specific predictability since both H_0^A and H_0^B are strongly rejected. We also note that inferences based on $\text{SupWald}^{B,\text{ivx}}$ appear robust to alternative choices of δ in the construction of the IVX variable. It is also interesting to note that unlike the linear case, inferences appear to be robust to the starting period. One should be cautious, however, when interpreting inferences such as the ones based on the 1990–2007 period due to sample size limitations, which are further exacerbated when fitting a threshold specification.

Recalling that the R^2's characterizing the various linear specifications were clustered around values close to zero (see Table 7), it is also useful to highlight the remarkable jump in goodness of fit in our proposed threshold model presented in (15). Our results strongly point toward the presence of very strong predictability during *bad times* when the growth in industrial production (IP) (variable $\Delta IP_t = \ln(IP_t/IP_{t-1})$) is negative, while no or very weak predictability during expansionary periods or normal times. More specifically, over the 1950–2007 period, we have

$$\hat{r}_{t+1} = \begin{cases} 0.1606_{(0.0357)} + 0.0441_{(0.0107)}LDY_t \\ \qquad \Delta IP_t \leq -0.0036,\ R_1^2 = 17.47\%,\ N_1 = 131 \\ 0.0135_{(0.0161)} + 0.0010_{(0.0045)}LDY_t \\ \qquad \Delta IP_t > -0.0036,\ R_2^2 = 0.00\%,\ N_2 = 564, \end{cases} \tag{15}$$

with a joint R^2 of 3.88%. Estimated standard errors are in parentheses. Besides being interesting in its own right, this result may also help explain the conflicting results obtained in the recent literature where the samples considered included or excluded data on the late 1990s and 2000s, a period with few recessions. Even with the reduction in the sample size, it is quite

Table 8. Regime-specific predictability

		$\text{SupWald}^{B,\text{ivx}}$		
	SupWald^A	$\delta = 0.7$	$\delta = 0.8$	$\delta = 0.9$
1950–2007	20.75 (0.001)	26.75*	28.87*	30.21*
1960–2007	18.98 (0.002)	23.24*	23.40*	23.46*
1970–2007	17.73 (0.004)	21.64*	21.82*	21.77*
1980–2007	24.52 (0.000)	27.73*	28.60*	28.96*
1990–2007	28.87 (0.000)	29.52*	30.18*	31.10*

remarkable that the goodness of fit can jump from a magnitude close to zero to about 17% in one subset. Overall, our results strongly support DY-based predictability in U.S. returns but occurring solely during *bad times*. Note, for instance, that more than half of the periods during which $\Delta \text{IP}_t \leq -0.0036$ coincide with the NBER (National Bureau of Economic Research) recessions. The strength of this predictability is very strong and unlikely to be sensitive to the methodology or our assumptions. Interestingly and through a different methodology, our findings about the presence of strong return predictability during bad times also corroborate the findings in Henkel et al. (2009). Using Bayesian inference techniques on a Markov switching VAR (vector autoregression) setup in which they consider multiple predictors in addition to the DY, the authors document a substantial jump in predictive strength of variables such as DY, short-term rates, term structure, etc., during recessions.

6. CONCLUSIONS

The goal of this article was to develop inference methods useful for detecting the presence of regime-specific predictability in predictive regressions. We obtained the limiting distributions of a series of Wald statistics designed to test the null of linearity versus threshold-type nonlinearity and the joint null of linearity and no predictability. One important feature of the limiting distribution that arises in the first case is the fact that it does not depend on any unknown nuisance parameters, thus making it straightforward to use. This is an unusual occurrence in this literature, where under a purely stationary framework (as opposed to a nearly integrated one), it is well known that limiting distributions typically depend on unknown population moments of the underlying models.

Our empirical application also leads to the interesting result that U.S. return series are clearly predictable using valuation ratios such as DY, but this predictability kicks in solely during bad times and would therefore be masked in studies that operate within linear specifications.

Finally, some important extensions to the present work are worth mentioning. A useful extension we are currently considering involves introducing long-horizon variables into (1)–(2). This would offer an interesting parallel to the linear predictive regression literature, which has often distinguished long- versus short-horizon predictability. Other important extensions include extending (1)–(2) to allow for more than two regimes, following some of the methods developed in Gonzalo and Pitarakis (2002), while further statistical properties (e.g., confidence intervals) of objects such as the estimated threshold parameter may be explored using the subsampling methodology of Gonzalo and Wolf (2005).

A key assumption under which we have operated ruled out heteroscedasticity and serial correlation in u_t. As our empirical application has documented however, our results can continue to be extremely useful despite this limitation. This restriction is in fact the norm rather than the exception in any work that introduced nonlinearities parametrically or nonparametrically in models that contain persistent variables. Albeit challenging, we expect future work to also be directed toward tackling these issues.

APPENDIX: PROOFS

Lemma 1. Under Assumptions 1 and 2 and as $T \to \infty$, we have (a) $\frac{\sum I_{1t}}{T} \xrightarrow{p} \lambda$, (b) $\frac{\sum x_t}{T^{\frac{3}{2}}} \Rightarrow \int_0^1 K_c(r)dr$, (c) $\frac{\sum x_t^2}{T^2} \Rightarrow \int_0^1 K_c^2(r)dr$, (d) $\frac{\sum x_{t-1}v_t}{T} \Rightarrow \int_0^1 K_c(r)dB_v(r) + \lambda_{vv}$. (e) $\frac{\sum x_{t-1}u_t}{T} \Rightarrow \int_0^1 K_c(r)dB_u(r,1)$, (f) $\frac{\sum x_t^2 I_{1t}}{T^2} \Rightarrow \lambda \int_0^1 K_c^2(r)dr$, (g) $\frac{\sum x_t I_{1t}}{T^{\frac{3}{2}}} \Rightarrow \lambda \int_0^1 K_c(r)dr$, (h) $\frac{\sum_{t=1}^{[Tr]} u_t I_{1t-1}}{\sqrt{T}} \Rightarrow B_u(r,\lambda)$, (i) $\frac{\sum x_{t-1}u_t I_{1t-1}}{T} \Rightarrow \int_0^1 K_c(r)dB_u(r,\lambda)$

Proof of Lemma 1. (a) By Assumptions 1 and 2, I_{1t} is strong mixing with the same mixing numbers as q_t. The result then follows from a suitable law of large numbers (see White 2001, secs. 3.3– 3.4). (b)–(e) Under our Assumptions 1 and 2, the results follow directly from lemma 3.1 in Phillips (1988). (f) Letting $X_{T,t} = x_t/\sqrt{T}$ and $X_T(r) = x_{[Tr]}/\sqrt{T}$, we can rewrite (f) as

$$\frac{1}{T}\sum X_{T,t}^2 I_{1t} = \lambda \frac{1}{T}\sum X_{T,t}^2 + \frac{1}{T}\sum X_{T,t}^2 (I_{1t} - \lambda). \quad \text{(A.1)}$$

Under Assumptions 1 and 2 and requiring $E|e_t|^p < \infty$ for some $p \geq 4$, we can make use of the strong approximation result $\sup_{r\in[0,1]} |X_T(r) - K_c(r)| = o_p(T^{-a})$, with $a = (p-2)/2p$ (see lemma A.3 in Phillips 1998, and Phillips and Magdalinos 2007), to obtain

$$\frac{1}{T}\sum X_{T,t}^2 = \int_0^1 K_c^2(r)dr + o_p(T^{-a}). \quad \text{(A.2)}$$

Indeed,

$$\begin{aligned}
&\left| \int_0^1 X_T(r)^2 dr - \int_0^1 K_c(r)^2 dr \right| \\
&\leq \int_0^1 \left| X_T(r)^2 - K_c(r)^2 \right| dr \\
&= \int_0^1 |X_T(r) - K_c(r)|\, |X_T(r) + K_c(r)|\, dr \\
&\leq \sup_r |X_T(r) - K_c(r)| \left(\sup_r |X_T(r)| + \sup_r |K_c(r)| \right) \\
&= o_p(T^{-a}). \quad \text{(A.3)}
\end{aligned}$$

The above then leads to

$$\begin{aligned}
&\frac{1}{T}\sum X_{T,t}^2 I_{1t} - \lambda \int_0^1 K_c(r)^2 dr \\
&\quad = \frac{1}{T}\sum X_{T,t}^2 (I_{1t} - \lambda) + o_p(T^{-a}), \quad \text{(A.4)}
\end{aligned}$$

holding uniformly $\forall \lambda \in \Lambda$. Finally, given that $\sup_{r\in[0,1]} |X_T(r)| = O_p(1)$, together with the fact that the result in (a) also holds uniformly over λ (see lemma 1 in Hansen 1996), we have $\sup_\lambda |\frac{1}{T}\sum X_{T,t}^2 I_{1t} - \lambda \int_0^1 K_c(r)^2 dr| = o_p(1)$, implying the required result. (g) follows identical lines to the proof of (f). (h)–(i) Since our assumptions satisfy their assumption 2, the result in (h) is theorem 1 of Caner and Hansen (2001), while our result in (i) follows along the same lines as theorem 2 of Caner and Hansen (2001).

Proof of Proposition 1. It is initially convenient to reformulate $W_T^A(\lambda)$ under H_0^A as

$$W_T^A(\lambda) = \left[u'X_1 - u'X(X'X)^{-1}X_1'X_1\right] \times \left[X_1'X_1 - X_1'X_1(X'X)^{-1}X_1'X_1\right]^{-1} \times \left[X_1'u - (X_1'X_1)(X'X)^{-1}X'u\right]/\hat{\sigma}_u^2, \quad \text{(A.5)}$$

where X_i is the matrix stacking $(I_{it}\ x_t I_{it})$ for $i = 1, 2$. With $D_T = \text{diag}(\sqrt{T}, T)$, we can write

$$D_T^{-1}X_1'X_1D_T^{-1} = \begin{pmatrix} \frac{\sum I_{1t}}{T} & \frac{\sum x_t I_{1t}}{T^{\frac{3}{2}}} \\ \frac{\sum x_t I_{1t}}{T^{\frac{3}{2}}} & \frac{\sum x_t^2 I_{1t}}{T^2} \end{pmatrix}, \quad \text{(A.6)}$$

and using Lemma 1, we have the following weak convergence results:

$$D_T^{-1}X_1{}'X_1D_T^{-1} \Rightarrow \begin{pmatrix} \lambda & \lambda\int_0^1 K_c(r)dr \\ \lambda\int_0^1 K_c(r)dr & \lambda\int_0^1 K_c^2(r)dr \end{pmatrix} \equiv \lambda\int_0^1 \overline{K}_c(r)\overline{K}_c(r)' \quad \text{(A.7)}$$

and

$$D_T^{-1}X'XD_T^{-1} \Rightarrow \int_0^1 \overline{K}_c(r)\overline{K}_c(r)', \quad \text{(A.8)}$$

where $\overline{K}_c(r) = (1, K_c(r))$. It now follows from the continuous mapping theorem (CMT) that

$$\left[D_T^{-1}X_1'X_1D_T^{-1} - D_T^{-1}X_1'X_1(X'X)^{-1}X_1'X_1D_T^{-1}\right]^{-1} \Rightarrow \frac{1}{\lambda(1-\lambda)}\left(\int_0^1 \overline{K}_c(r)\overline{K}_c(r)'\right)^{-1}. \quad \text{(A.9)}$$

We next focus on the limiting behavior of $D_T^{-1}X'u$ and $D_T^{-1}X_1'u$. Looking at each component separately, setting $\sigma_u^2 = 1$ for simplicity and no loss of generality and using Lemma 1, we have

$$D_T^{-1}X_1'u = \begin{pmatrix} \frac{\sum I_{1t}u_{t+1}}{\sqrt{T}} \\ \frac{\sum x_t I_{1t}u_{t+1}}{T} \end{pmatrix} \Rightarrow \begin{pmatrix} B_u(r,\lambda) \\ \int_0^1 K_c(r)dB_u(r,\lambda) \end{pmatrix}, \quad \text{(A.10)}$$

and

$$D_T^{-1}X'u = \begin{pmatrix} \frac{\sum u_{t+1}}{\sqrt{T}} \\ \frac{\sum x_t u_{t+1}}{T} \end{pmatrix} \Rightarrow \begin{pmatrix} B_u(r,1) \\ \int_0^1 K_c(r)dB_u(r,1) \end{pmatrix}. \quad \text{(A.11)}$$

The above now allows us to formulate the limiting behavior of $D_T^{-1}X_1'u - \lambda D_T^{-1}X'u$ as

$$D_T^{-1}X_1'u - \lambda D_T^{-1}X'u \Rightarrow \int_0^1 \overline{K}_c(r)dG_u(r,\lambda), \quad \text{(A.12)}$$

where $G_u(r,\lambda) = B_u(r,\lambda) - \lambda B_u(r,1)$. The result in (3) follows straightforwardly through the use of the CMT and standard algebra.

Proof of Proposition 2. We rewrite our most general unrestricted specification in (1) as $y = \alpha_1 I_1 + \beta_1 x_1 + \alpha_2 I_2 + \beta_2 x_2 + u$. Within this notation, lower-case $x_i's$ stack $x_t I_{it}$, while the $I_i's$ stack I_{it} for $i = 1, 2$. We also recall that $X_i = (I_i\ x_i)$ for $i = 1, 2$. It is now convenient to reformulate (1) as $y = \alpha + \beta x + X_2\eta + u$, where $\alpha = \alpha_1$, $\beta = \beta_1$, and $\eta = (\gamma, \delta)'$, with $\gamma = \alpha_2 - \alpha_1$ and $\delta = \beta_2 - \beta_1$ so that within this alternative parameterization, $H_0^A : \eta = 0$ and $H_0^B : \eta = 0, \beta = 0$. Next, consider a most general (MG) model containing $(1\ \ x\ \ X_2) = (X\ \ X_2)$, a partially restricted (PR) version containing $X = (1\ \ x)$, and a fully restricted (FR) version containing the vector of 1's. From standard projection algebra, the sum of squared errors corresponding to each specification are $SSE_{MG} = y'M_{X,X_2}y$, $SSE_{PR} = y'M_Xy$, and $SSE_{FR} = y'M_1y$, where $M_X = I - X(X'X)^{-1}X'$ and $M_{X,X_2} = M_X - M_XX_2(X_2'M_XX_2)^{-1}X_2'M_X$. It now trivially follows that we can write the Wald statistics corresponding to each hypothesis as $W_T^A(\lambda) = [y'M_Xy - y'M_{X,X_2}y]/\hat{\sigma}_u^2$ (PR against MG), $W_T^B(\lambda) = [y'M_1y - y'M_{X,X_2}y]/\hat{\sigma}_u^2$ (FR against MG), and $W_T(\beta = 0) = [y'M_1y - y'M_Xy]/\hat{\sigma}_{\text{lin}}^2$ (FR against PR). It can now immediately be observed that $W_T^B(\lambda) = W_T^A(\lambda) + (\hat{\sigma}_{\text{lin}}^2/\hat{\sigma}_u^2)W_T(\beta = 0)$. Under the null hypothesis, $(\hat{\sigma}_{\text{lin}}^2/\hat{\sigma}_u^2) \xrightarrow{p} 1$ and therefore in large samples, $W_T^B(\lambda) \approx W_T(\beta = 0) + W_T^A(\lambda)$ and $\sup_\lambda W_T^B(\lambda) \approx W_T(\beta = 0) + \sup_\lambda W_T^A(\lambda)$, as required. To obtain the limiting distribution in (5), it now suffices to use the results presented in Lemma 1, together with the CMT along lines identical to those in the proof of Proposition 1.

Proof of Proposition 3. Our result in (13) follows directly from (11)–(12), theorem 3.8 in PM09 (p. 14), Lemma 1, Proposition 1, and the use of the CMT. Note that theorem 3.8 in PM09 has been obtained within a model with no fitted intercept; however, Stamatogiannis (2010, theorem 4.2, p. 154) and Kostakis et al. (2010) also established its validity in the more general setting that includes a constant term and a predictive regression setting identical to our specification in (7) and thus leading to our own result.

SUPPLEMENTARY MATERIALS

Appendix: File providing additional Monte Carlo simulations and further details on some of the proofs.

ACKNOWLEDGMENTS

J. Gonzalo thanks the Spanish Ministerio de Ciencia e Innovacion, grants SEJ2007-63098, ECO2010-19357, CONSOLIDER 2010 (CSD 2006-00016), DGUCM (Community of Madrid) grant EXCELECON S-2007/HUM-044 and Bank of Spain (ER program) for partially supporting this research. J. Pitarakis thanks the ESRC for partially supporting this research through an individual research grant RES-000-22-3983. Both authors are grateful to Grant Hillier, Tassos Magdalinos, Peter Phillips, and Peter Robinson for very useful suggestions and helpful discussions. Detailed comments and feedback from the editor, associate editor and two anonymous referees are also gratefully acknowledged. Last but not least, we also thank seminar participants at Queen-Mary, LSE, Southampton, Exeter, Manchester, and Nottingham universities, the ESEM 2009 meetings in

Barcelona, the SNDE 2010 meeting in Novara, and the 2010 CFE conference in London for useful comments. All errors are our own responsibility.

REFERENCES

Andrews, D. W. K. (1993), "Tests for Parameter Instability and Structural Change With Unknown Change Point," *Econometrica*, 61, 821–856. [232,234]

Ang, A., and Bekaert, G. (2007), "Stock Return Predictability: Is It There?," *Review of Financial Studies*, 20, 651–707. [230,237]

Bandi, F., and Perron, B. (2008), "Long-Run Risk-Return Trade-Offs," *Journal of Econometrics*, 143, 349–374. [229]

Campbell, J. Y., and Yogo, M. (2006), "Efficient Tests of Stock Return Predictability," *Journal of Financial Economics*, 81, 27–60. [229,230,231,237,237]

Caner, M., and Hansen, B. E. (2001), "Threshold Autoregression With a Unit Root," *Econometrica*, 69, 1555–1596. [230,232,239]

Cavanagh, C. L., Elliott, G., and Stock, J. H. (1995), "Inference in Models With Nearly Integrated Regressors," *Econometric Theory*, 11, 1131–1147. [230,231]

Chan, N. (1988), "The Parameter Inference for Nearly Nonstationary Time Series the Parameter Inference for Nearly Nonstationary Time Series," *Journal of the American Statistical Association*, 83, 857–862. [230]

Cochrane, J. H. (2008), "The Dog That Did Not Bark: A Defense of Return Predictability," *Review of Financial Studies*, 21, 1533–1575. [229,237]

Elliott, G. (1998), "On the Robustness of Cointegration Methods When Regressors Almost Have Unit Roots," *Econometrica*, 66, 149–158. [233]

Estrella, A. (2003), "Critical Values and P Values of Bessel Process Distributions: Computation and Application to Structural Break Tests," *Econometric Theory*, 19, 1128–1143. [232,234]

Gonzalo, J., and Pitarakis, J. (2002), "Estimation and Model Selection Based Inference in Single and Multiple Threshold Models," *Journal of Econometrics*, 110, 319–352. [239]

Gonzalo, J., and Wolf, M. (2005), "Subsampling Inference in Threshold Autoregressive Models," *Journal of Econometrics*, 127, 201–224. [239]

Hansen, B. E. (1996), "Inference When a Nuisance Parameter Is Not Identified Under the Null Hypothesis," *Econometrica*, 64, 413–430. [232,235,239]

——— (1997), "Approximate Asymptotic P-Values for Structural Change Tests," *Journal of Business and Economic Statistics*, 15, 60–67. [232,234]

——— (1999), "Testing for Linearity," *Journal of Economic Surveys*, 13, 551–576. [235]

Henkel, S. J., Martin, J. S., and Nardari, F. (2009), "Time-Varying Short-Horizon Return Predictability," AFA 2008 New Orleans Meetings Paper [on-line]. Available at *http://ssrn.com/abstract=1101944*. [237,239]

Jansson, M., and Moreira, M. J. (2006), "Optimal Inference in Regression Models With Nearly Integrated Regressors," *Econometrica*, 74, 681–714. [229,230,231,237]

Kostakis, A., Magdalinos, A., and Stamatogiannis, M. (2010), "Econometric Inference for Stock Return Predictability," unpublished manuscript, University of Nottingham, UK. [233,240]

Lettau, M., and Van Nieuwerburgh, S. (2008), "Reconciling the Return Predictability Evidence," *Review of Financial Studies*, 21, 1607–1652. [229,237]

Lewellen, J. (2004), "Predicting Returns With Financial Rations," *Journal of Financial Economics*, 74, 209–235. [229,237,237]

Menzly, L., Santos, T., and Veronesi, P. (2004), "Understanding Predictability," *Journal of Political Economy*, 112, 1–47. [237]

Park, J. Y., and Phillips, P. C. B. (1988), "Statistical Inference in Regressions With Integrated Processes: Part 1," *Econometric Theory*, 4, 468–497. [232]

Petruccelli, J. D. (1992), "On The Approximation of Time Series by Threshold Autoregressive Models," *Sankhya,* Series B, 54, 54–61. [230]

Phillips, P. C. B. (1988), "Regression Theory for Near-Integrated Time Series," *Econometrica*, 56, 1021–1043. [231,239]

——— (1998), "New Tools for Understanding Spurious Regressions," *Econometrica*, 66, 1299–1325. [239]

Phillips, P. C. B., and Hansen, B. E. (1990), "Statistical Inference in Instrumental Variables Regression With I(1) Process," *Review of Economic Studies*, 57, 99–125. [233]

Phillips, P. C. B., and Magdalinos, A. (2007), "Limit Theory for Moderate Deviations From a Unit Root Under Weak Dependence," *Journal of Econometrics*, 136, 115–130. [239]

——— (2009), "Econometric Inference in the Vicinity of Unity," CoFie Working Paper No. 7, Singapore Management University, Singapore. [233]

Pitarakis, J. (2008), "Threshold Autoregressions With a Unit Root: Comment," *Econometrica*, 76, 1207–1217. [230,232]

Rapach, D. E., and Wohar, M. E. (2006), "Structural Breaks and Predictive Regression Models of Aggregate U.S. Stock Returns," *Journal of Financial Econometrics*, 4, 238–274. [237]

Rossi, B. (2005), "Optimal Tests for Nested Model Selection With Underlying Parameter Instability," *Econometric Theory*, 21, 962–990. [233,237]

——— (2006), "Are Exchange Rates Really Random Walks? Some Evidence Robust to Parameter Instability," *Macroeconomic Dynamics*, 10, 20–38. [237]

Saikkonen, P. (1991), "Asymptotically Efficient Estimation of Cointegrating Regressions," *Econometric Theory*, 7, 1–21. [233]

——— (1992), "Estimation and Testing of Cointegrated Systems by an Autoregressive Approximation," *Econometric Theory*, 8, 1–27. [233]

Stamatogiannis, M. P. (2010), "Econometric Inference in Models With Nonstationary Time Series," Ph.D. Dissertation, Department of Economics, University of Nottingham. Available at *http://etheses.nottingham.ac.uk/1950/1/StamatogiannisThesis2010.pdf* [233,240]

Timmermann, A. (2008), "Elusive Return Predictability," *International Journal of Forecasting*, 24, 1–18. [237]

Valkanov, R. (2003), "Long-Horizon Regressions: Theoretical Results and Applications," *Journal of Financial Economics*, 68, 201–232. [229,237]

Van der Vaart, J., and Wellner, J. (1996), *Weak Convergence and Empirical Processes: With Applications to Statistics*, New York: Springer-Verlag. [235]

White, H. (2001), *Asymptotic Theory for Econometricians* (2nd rev. ed.), New York: Academic Press. [239]